MM-LEJEPA: CORRELATED GAUSSIAN REGULARIZATION FOR MULTIMODAL JOINT-EMBEDDING PREDICTIVE ARCHITECTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

LeJEPA derives an isotropic Gaussian embedding target by minimizing, under its task assumptions, a Fisher information criterion that controls the task-averaged squared prediction bias. This enables unimodal self-supervised learning without reconstruction decoders, exponential-moving-average teachers, or contrastive negatives. However, its isotropic task prior does not account for the cross-modal structure of multimodal tasks that read shared content. We introduce MM-LeJEPA by extending LeJEPA's task assumptions to multimodal learning. Specifically, we model multimodal tasks as strength-weighted combinations of components that read independent information from the two modalities and components that read the same content from both. Under explicit gradient-moment, correspondence, and symmetry assumptions, we derive a correlated Gaussian target that minimizes a weighted Fisher information functional under a fixed total variance constraint. An analytic transform adapts Sketched Isotropic Gaussian Regularization (SIGReg) to this target, yielding a two-term objective that combines distributional regularization with local-to-global prediction. The extension adds a single scalar hyperparameter, the target cross-modal correlation, and retains the simplicity of LeJEPA. On public audio-visual classification benchmarks, our results are competitive with representative open-source methods. Code will be released.

1 INTRODUCTION

Self-supervised learning aims to learn representations from unlabeled data that transfer to unknown downstream tasks. Across unimodal and multimodal settings, common approaches include contrastive learning (Chen et al., 2020b; He et al., 2020), self-distillation (Grill et al., 2020; Caron et al., 2021), and masked reconstruction (He et al., 2022; Georgescu et al., 2023). Audio-visual methods combine these strategies with cross-modal alignment (Gong et al., 2023; Huang et al., 2023), fine-grained synchronization (Araujo et al., 2025), and dedicated fusion mechanisms (Nagrani et al., 2021; Li et al., 2023; 2024). While effective, these designs require choices about masking, alignment losses, and encoder sharing, without explicitly deriving a target joint embedding distribution from a prior over downstream tasks. This design and tuning burden makes it harder to develop a simple, reusable multimodal pretraining framework.

Joint-embedding predictive architectures (JEPAs) provide a self-supervised learning framework that predicts one view's representation from another rather than reconstructing inputs (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024). This removes the need for an input-reconstruction decoder, but a collapse-prevention mechanism is still required (Assran et al., 2023; Balestriero & LeCun, 2025; Bardes et al., 2022). LeJEPA (Balestriero & LeCun, 2025) instead asks what distribution the embeddings should follow. It shows that, when nothing is known in advance about the downstream task and the total variance is fixed, an isotropic Gaussian, which spreads equally in all directions, minimizes a criterion that controls the average prediction bias of downstream probes. It then uses a lightweight regularizer, Sketched Isotropic Gaussian Regularization (SIGReg), to pull the embeddings toward this distribution, which enables efficient representation learning without a reconstruction decoder, contrastive negatives, stop-gradient, or an exponential-moving-average (EMA) teacher.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

LeJEPA was developed for a single modality, and its task prior treats all embedding directions alike. Multimodal learning, however, differs from unimodal learning in the structure of downstream tasks. With one encoder per modality, the joint embedding splits into an audio block and a video block. Audio and video may observe the same event, so part of the content captured by one encoder is also captured by the other, and a downstream task can read this shared content from both blocks at once. Applying LeJEPA's isotropic Gaussian target directly to the concatenated audio-visual embedding penalizes exactly the cross-modal correlation that shared content induces, limiting the target's suitability for tasks that read such content.

We therefore propose MM-LeJEPA, a multimodal extension of LeJEPA. The formulation applies to any pair of modalities; we develop and evaluate it for audio and video, the most common co-occurring pair. We single out a class of tasks specific to multimodal learning, those that read shared content from both modalities, and model downstream tasks as strength-weighted combinations of independent-information and shared-content components. Under assumptions that, like LeJEPA's, favor no feature direction within each modality, this yields a task prior suited to multimodal learning. The optimal embedding distribution under this prior is a correlated Gaussian: each modality remains an isotropic Gaussian, while the audio and video embeddings are correlated through one coefficient ρ. The target reduces to LeJEPA's isotropic Gaussian when shared content is absent, and its correlation grows with the contribution of shared content. Our contributions are:

- MM-LeJEPA extends LeJEPA to multimodal learning by accounting for both independent and shared information in its task prior. The objective retains two terms, and the distributional extension introduces only one target-correlation parameter ρ, without requiring reconstruction decoders, EMA teachers, or contrastive negatives.
- We model independent-information and shared-content task components and derive a correlated Gaussian that minimizes the resulting weighted Fisher information functional under explicit assumptions and a total-variance constraint. A fixed whitening transform makes this target compatible with SIGReg, and local-to-global prediction preserves content.
- Comparisons with open-source methods on public audio-visual benchmarks show competitive attentive-probing results and fine-tuning performance.

2 Related Work

2.1 Audio-visual self-supervised pretraining

Paired audio and video provide natural self-supervision: early work transferred visual knowledge to audio (Aytar et al., 2016), predicted audio-image correspondence (Arandjelović & Zisserman, 2017; 2018) or temporal synchronization (Korbar et al., 2018; Owens & Efros, 2018), and later work added cross-view contrastive learning (Tian et al., 2020), cross-modal clustering supervision (Alwassel et al., 2020) and refined positive pairs (Morgado et al., 2021b;a; Patrick et al., 2021). Recent methods reconstruct masked audio-visual content: Audiovisual MAE reconstructs both modalities with a shared decoder, trained from scratch on the target dataset (Georgescu et al., 2023); CAV-MAE and MAViL add cross-modal contrastive learning, and MAViL also adds a self-training stage with targets from a teacher that sees the complete input (Gong et al., 2023; Huang et al., 2023); CAV-MAE Sync aligns individual audio segments with video frames (Araujo et al., 2025). Others vary the architecture or enforce augmentation consistency: MBT fuses modalities through a few bottleneck tokens (Nagrani et al., 2021), AVSiam shares one Vision Transformer across modalities (Lin & Bertasius, 2024), and EquiAV models equivariance to input augmentations (Kim et al., 2024).

These methods encode cross-modal relationships through reconstruction, alignment or architectural choices, without deriving a target joint embedding distribution from a prior over downstream tasks. MM-LeJEPA specifies such a distribution and exposes its cross-modal correlation as a scalar ρ, and needs no reconstruction decoder, contrastive negatives, EMA teacher or self-training stage.

2.2 Joint-embedding predictive architectures and LeJEPA

JEPAs predict one view's embedding from another's instead of reconstructing inputs (LeCun, 2022), so the encoder can discard unpredictable low-level detail. I-JEPA applies this to image blocks (Assran et al., 2023), V-JEPA and V-JEPA 2 to video (Bardes et al., 2024; Assran et al., 2025), A-JEPA

to audio (Fei et al., 2023), data2vec to speech, vision and language (Baevski et al., 2022), and iBOT learns an online tokenizer by self-distillation (Zhou et al., 2022). Since latent prediction admits the constant solution, these methods prevent collapse with stop-gradient and a moving-average teacher, as does MJEPA for intra- and cross-modal audio-visual prediction with a modality-shared encoder (Teotia et al., 2026). A parallel line uses explicit geometric constraints instead: contrastive repulsion (Chen et al., 2020b; He et al., 2020), cross-correlation or variance–covariance penalties (Zbontar et al., 2021; Bardes et al., 2022), alignment and uniformity on the hypersphere (Wang & Isola, 2020), and batch-estimated whitening (Ermolov et al., 2021).

LeJEPA prevents collapse by using SIGReg to regularize embeddings toward an isotropic Gaussian target (Balestriero & LeCun, 2025). This target follows from its analysis of task-averaged integrated squared bias under a task-gradient prior and regularity assumptions. LeVJEPA extends global–local invariance with SIGReg to video (Kuhn et al., 2026), AV-JEPA combines early fusion and modality dropout with the LeJEPA objective without an EMA teacher (Robson et al., 2026), and Le MuMo JEPA pairs RGB images with depth or thermal images through learnable fusion tokens while keeping the isotropic target (Cornelissen et al., 2026).

Rather than applying LeJEPA's isotropic target unchanged, MM-LeJEPA derives a joint target from explicit assumptions about multimodal tasks: a shared-content component adds cross-modal structure to the second moment of the task gradient, and minimizing the resulting weighted Fisher information functional under a fixed total variance yields a correlated Gaussian target.

3 METHOD

Overview. We follow LeJEPA's distribution-design approach. Section 3.1 expresses it in a matrix-weighted form: a prior over downstream tasks fixes the second moment W of the task gradient and yields the weighted Fisher functional J_W that controls the bias of nonlinear probes, and minimizing J_W under a fixed total variance yields the target embedding distribution. Section 3.2 derives W from a multimodal task prior with a shared-content component, Section 3.3 minimizes the corresponding functional $J_{W_{\mathrm{AV}}}$ and obtains a correlated Gaussian target with a single parameter ρ, and Section 3.4 enforces this target with SIGReg through a fixed whitening transform.

3.1 PRELIMINARIES: LEJEPA'S ASSUMPTIONS ABOUT DOWNSTREAM TASKS

Let $\mathbf{Z} \in \mathbb{R}^{N\times K}$ denote the matrix of N embeddings produced by the encoder f_θ from the inputs $\mathbf{x}_n$, where each row is a K-dimensional embedding $f_\theta(\mathbf{x}_n)^\top$. We write p_z for the distribution of a single embedding $\mathbf{z} = f_\theta(\mathbf{x})$. As in LeJEPA, p_z has zero mean and covariance $\Sigma \succ 0$. The unknown labels of a downstream task are given by a target function $\eta : \mathbb{R}^K \to \mathbb{R}$, with gradient $\nabla\eta(\mathbf{q})$ at a query point $\mathbf{q}$. This gradient describes the task's local sensitivity along each embedding direction.

LeJEPA's objective. We express LeJEPA's probe-bias analysis in a matrix-weighted form for task-averaged integrated squared bias (ISB). The analysis yields a quantity that depends on the embedding distribution and controls the ISB; minimizing it under the variance constraint gives the optimal distribution for this criterion. We first obtain this quantity, J_W.

We first consider a task prior fixed independently of the candidate representation distributions, and assume the task gradient has a zero first moment and a common positive-definite second moment:

$$\mathbb{E}_\eta[\nabla\eta(\mathbf{q})] = 0, \qquad \mathbb{E}_\eta\big[\nabla\eta(\mathbf{q})\nabla\eta(\mathbf{q})^\top\big] = W \succ 0. \tag{1}$$

Both conditions hold with the same W for almost every $\mathbf{q}$, with finite second moments; a zero mean indicates no preferred signed gradient direction, and the form of W is left unspecified.

Let $\mathcal{I}(p_z) := \mathbb{E}_{\mathbf{z}\sim p_z}\big[\nabla \log p_z(\mathbf{z})\, \nabla \log p_z(\mathbf{z})^\top\big]$ be the Fisher information matrix. We use the weighted Fisher information functional below as a surrogate objective, with weights given by the task-gradient second moment W; the second equality is proved in Appendix D.1:

$$J_W(p_z) := \int_{\mathbb{R}^K} (\nabla \log p_z(\mathbf{z}))^\top W\, (\nabla \log p_z(\mathbf{z}))\, p_z(\mathbf{z})\, \mathrm{d}\mathbf{z} = \mathrm{tr}(W\mathcal{I}(p_z))\,. \tag{2}$$

Following LeJEPA's task setting and derivations, we consider two nonlinear probes: radius-based k-NN (Cover & Hart, 1967) and the Nadaraya–Watson kernel probe (Nadaraya, 1964; Watson, 1964).

Under the task-gradient and curvature moment assumptions and each probe's regularity conditions, this functional controls their task-averaged ISB over queries $\mathbf{q} \sim p_z$:

$$\mathrm{ISB}_{k\mathrm{NN}} = c_{k\mathrm{NN},0} + c_{k\mathrm{NN},1}\, J_W(p_z) + \varepsilon_{k\mathrm{NN}}, \qquad \mathrm{ISB}_{\mathrm{ker}} \le c_{\mathrm{ker},0} + c_{\mathrm{ker},1}\, J_W(p_z) + \varepsilon_{\mathrm{ker}}. \quad (3)$$

Both relations are derived in Appendix D.2. Each c is a nonnegative constant independent of the candidate distribution's shape, and each ε is an infinitesimal term; the distribution-dependent design term is $J_W(p_z)$.

Optimizing the control quantity J_W. Following LeJEPA's distribution-design strategy, we minimize $J_W(p_z)$, which controls the task-averaged ISB, to obtain the optimal shape of the embedding distribution.

LeJEPA further assumes an isotropic second moment to model unknown downstream tasks without favoring any gradient direction. That is, W in Eq. (1) is taken to be $W = \tau_g^2 I_K$ with $0 < \tau_g^2 < \infty$.

Under this isotropic prior, Eq. (2) reduces to $J_W(p_z) = \tau_g^2 \operatorname{tr}\mathcal{I}(p_z) = \tau_g^2 J(p_z)$, where $J(p_z) := \operatorname{tr}\mathcal{I}(p_z) = \int_{\mathbb{R}^K} \|\nabla \log p_z(\mathbf{z})\|^2 p_z(\mathbf{z})\, \mathrm{d}\mathbf{z}$ is the Fisher information functional of LeJEPA. LeJEPA optimizes over all distributions of $\mathbf{z}$ subject to a fixed total-variance constraint. Taking that constraint to be $\operatorname{tr}(\Sigma) = K$, *i.e.* unit average coordinate variance,

$$J_W(p_z) = \tau_g^2 J(p_z) \;\ge\; \tau_g^2 \operatorname{tr}\big(\Sigma^{-1}\big) \;\ge\; \tau_g^2 K. \quad (4)$$

The two inequalities hold with equality if and only if $p_z = \mathcal{N}(0, \Sigma)$ and $\Sigma = I_K$, respectively; proof references for both inequalities and their equality conditions are given in Appendix D.3. Therefore, when $J_W(p_z)$ attains its minimum, p_z must satisfy both conditions, *i.e.*, $p_z^* = \mathcal{N}(0, I_K)$.

The constraint in LeJEPA. LeJEPA then constrains the embeddings with SIGReg, which drives the distribution to the standard Gaussian by matching random one-dimensional projections with a univariate normality test (Epps & Pulley, 1983). A single weight combines this distributional term with a prediction loss that pulls an input's crop embeddings toward the mean embedding of its global crops, similar to the local-to-global correspondence in multi-crop training (Caron et al., 2020; 2021). The resulting objective needs no reconstruction decoder, EMA teacher, or contrastive negatives: prediction preserves content, while SIGReg constrains the distribution.

Summary. The target is thus governed by the second moment W of the task prior: an isotropic W gives an isotropic Gaussian, and a prior that couples two modalities changes the target only through W, which the next subsections derive for multimodal tasks.

3.2 Multimodal task prior and the second moment of the joint task gradient

Figure 1 summarizes MM-LeJEPA from view construction to prediction and regularization.

From unimodal to multimodal tasks. Section 3.1 shows that the target is fixed by the second moment W of the task gradient, and LeJEPA takes W isotropic because nothing is known about the task. With two modalities this choice is no longer neutral. The coordinates of the joint embedding come from two encoders, and audio and video often observe the same event, so a task that reads such shared content has gradients on the two sides that vary together; its second moment has cross-modal blocks that an isotropic W sets to zero. LeJEPA's isotropic target would therefore penalize exactly this correlation. We keep LeJEPA's route, a task prior that fixes W and the induced functional J_W, and change only the prior so that W carries the cross-modal structure of shared-content tasks.

Multimodal task prior. As in LeJEPA, we minimize J_W to control the ISB of Eq. (3). This section constructs the task-prior second moment W_{AV} for multimodal tasks. Multimodal tasks differ from unimodal ones in that some read the same information from both modalities at once.

For an audio-visual task, let the projected audio and video embeddings be $\mathbf{e}_a, \mathbf{e}_v \in \mathbb{R}^d$ and their joint embedding be $\mathbf{z} = [\mathbf{e}_a; \mathbf{e}_v] \in \mathbb{R}^{2d}$. At a fixed query point $\mathbf{q}$, we split the task gradient into blocks with respect to $\mathbf{e}_a$ and $\mathbf{e}_v$:

$$\nabla\eta(\mathbf{q}) = \big[\nabla_a\eta^\top,\ \nabla_v\eta^\top\big]^\top, \qquad \nabla_a\eta, \nabla_v\eta \in \mathbb{R}^d. \quad (5)$$

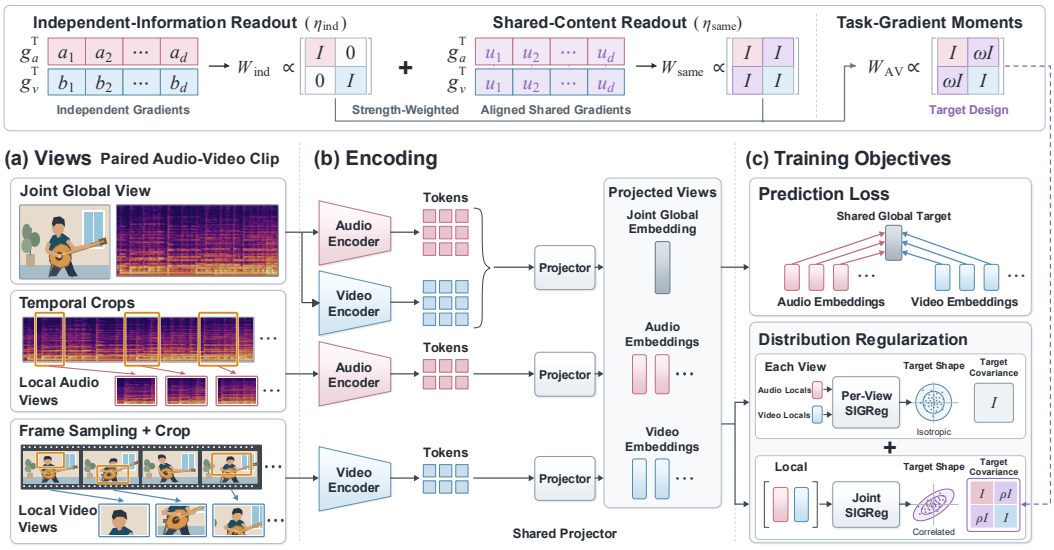

Figure 1: Overview of MM-LeJEPA. Top: the independent-information and shared-content readouts give the task-gradient second moments W_{ind} and W_{same}, whose weighted combination W_{AV} determines the correlated joint target. Audio crops and video frames give local views alongside a joint global view. Modality encoders and a shared projector produce embeddings for local-to-global prediction and distributional regularization with isotropic per-view and correlated joint targets.

Audio and video can provide different observations of the same content. We model a downstream task through independent-information and shared-content components, combining their gradient contributions with different strengths:

$$\nabla\eta = \sqrt{1-\pi}\,\nabla\eta_{\text{ind}} + \sqrt{\pi}\,\nabla\eta_{\text{same}}, \qquad 0 \le \pi < 1. \tag{6}$$

Here η_{ind} reads independent information from the two modalities and η_{same} reads the same content from both; the task prior is their combination, and π is a normalized squared-amplitude weight.

For example, identifying a guitar from its sound and image reads shared content, whereas recognizing background speech from audio and the shirt color from video reads independent information.

Fixed correspondence of multimodal content. The same content can be read out from both audio and video, but the directions and scales in which the two encoders represent it are arbitrary, and this arbitrariness carries over to the task distribution through the relation between the audio and video gradients of a shared-content task. To derive a target from the task prior, we therefore assume that this relation is fixed by how the two encoders represent the content and does not depend on the task; a single change of coordinates then aligns the two sides so that the gradient of shared content is the same in both modalities, and the task prior is formulated in terms of this fixed relation.

Assumption 1 (fixed correspondence of shared content)**.** *There exist a fixed orthogonal $R \in O(d)$ and a fixed positive diagonal $\Lambda \succ 0$, independent of the task and the query point, such that*

$$\nabla_v \eta_{\text{same}} = \Lambda R\, \nabla_a \eta_{\text{same}}. \tag{7}$$

Change of coordinates for a simpler form. An invertible change of coordinates leaves every task readout unchanged and maps each candidate embedding distribution to exactly one distribution in the new coordinates. We formulate the distribution-design problem in the common coordinates defined below. In Assumption 1, R describes the directional correspondence of shared-content readouts, while Λ accounts for their coordinate-wise scaling. We absorb these fixed differences into aligned video coordinates $\widetilde{\mathbf{e}}_v = R^\top \Lambda \mathbf{e}_v$. The same shared-content readout, expressed in $(\mathbf{e}_a, \widetilde{\mathbf{e}}_v)$, then satisfies

$$\nabla_{\widetilde{\mathbf{e}}_v} \eta_{\text{same}} = \nabla_a \eta_{\text{same}}. \tag{8}$$

All subsequent moments, probe calculations and variance constraints are defined in these common coordinates, and we henceforth write $\mathbf{e}_v$ for the aligned video coordinates; no per-task gradient correspondence is required for the independent-information component.

General assumptions on multimodal tasks. As in LeJEPA, the downstream task is not known in advance, so we place no preference on the direction of the task gradient: in the common coordinates, the gradient of each component has zero mean and an isotropic second moment within each modality, with the same strength on the two sides so that no modality is preferred, and the two components read mutually independent information. The shared-content component adds cross-modal structure under the correspondence and independence assumptions.

Assumption 2 (moments and independence of task gradients)**.** *For $m, m' \in \{a, v\}$, the independent-information component has independent modality gradients with*

$$\mathbb{E}_\eta[\nabla_m \eta_{\text{ind}}] = 0, \qquad \mathbb{E}_\eta[\nabla_m \eta_{\text{ind}} \nabla_{m'} \eta_{\text{ind}}^\top] = \delta_{mm'} \tau_{\text{ind}}^2 I_d, \tag{9}$$

where $\delta_{mm'}$ is the Kronecker delta. The shared-content component has gradients that coincide in the common coordinates by Eq. (8), with

$$\mathbb{E}_\eta[\nabla_m \eta_{\text{same}}] = 0, \qquad \mathbb{E}_\eta[\nabla_m \eta_{\text{same}} \nabla_{m'} \eta_{\text{same}}^\top] = \tau_{\text{same}}^2 I_d. \tag{10}$$

The two components are independent, hence $\mathbb{E}_\eta[\nabla \eta_{\text{ind}} \nabla \eta_{\text{same}}^\top] = 0$. Here $0 < \tau_{\text{ind}}^2, \tau_{\text{same}}^2 < \infty$.

Under the assumptions above, let $W_{\text{AV}} := \mathbb{E}_\eta[\nabla\eta\nabla\eta^\top]$ be the joint task-gradient second moment.

Proposition 1. *Under Assumptions 1 and 2, the component gradient second moments are*

$$W_{\text{ind}} = \tau_{\text{ind}}^2 \begin{bmatrix} I_d & 0 \\ 0 & I_d \end{bmatrix}, \qquad W_{\text{same}} = \tau_{\text{same}}^2 \begin{bmatrix} I_d & I_d \\ I_d & I_d \end{bmatrix}, \tag{11}$$

so that the second moment of the joint task gradient is

$$W_{\text{AV}} = (1-\pi) W_{\text{ind}} + \pi W_{\text{same}} = \tau_{\text{AV}}^2 \begin{bmatrix} I_d & \omega I_d \\ \omega I_d & I_d \end{bmatrix}, \tag{12}$$

with $\tau_{\text{AV}}^2 = (1-\pi)\tau_{\text{ind}}^2 + \pi\tau_{\text{same}}^2 > 0$ and $\omega = \pi\tau_{\text{same}}^2 / \tau_{\text{AV}}^2 \in [0, 1)$.

The second-moment expression is derived in Appendix D.4. At $\omega = 0$, $W_{\text{AV}} = \tau_{\text{AV}}^2 I_{2d}$ recovers LeJEPA's isotropic second-moment assumption over the full embedding space in Section 3.1. This isotropic prior ignores cross-modal coupling, whereas ω controls the contribution of shared-content components to the gradient second moment.

Summary. This section derived the task prior for multimodal tasks and its second moment W_{AV}. As in LeJEPA, the next section minimizes the Fisher information functional $J_{W_{\text{AV}}}$ induced by this prior under the total-variance constraint to obtain the optimal embedding distribution.

3.3 The optimal joint target distribution

With W_{AV} in hand, this section solves for the optimal distribution under the multimodal prior: among all embedding distributions with a fixed total variance, find the one that minimizes $J_{W_{\text{AV}}}$, the quantity that controls the probes' bias.

Optimizing the control quantity $J_{W_{\text{AV}}}$. With W_{AV} from Proposition 1, the weighted Fisher information functional of Eq. (2) reads $J_{W_{\text{AV}}}(p_z) = \text{tr}(W_{\text{AV}} \mathcal{I}(p_z))$; we minimize it under a fixed total-variance constraint to obtain the optimal shape of the embedding distribution.

Proposition 2. *Let p_z be an everywhere-positive density on $\mathbb{R}^{2d}$ with zero mean and covariance $\Sigma \succ 0$. Assume $p_z \in C^3$ with bounded derivatives up to order three, and assume finite Fisher information and covariance. Then*

$$J_{W_{\text{AV}}}(p_z) \;\geq\; \text{tr}\big(W_{\text{AV}} \Sigma^{-1}\big), \tag{13}$$

and, when $W_{\text{AV}} \succ 0$, equality holds if and only if $p_z = \mathcal{N}(0, \Sigma)$. The proof is given in Appendix D.5.

As in LeJEPA, we impose the total-variance constraint $\text{tr}(\Sigma) = 2d$, which only fixes the overall scale; the optimal covariance then takes the form below, with unit variance per coordinate.

Proposition 3. *Let W_{AV} have the form in Eq. (12), with $\omega \in [0, 1)$. Under the total-variance constraint above, the unique minimizer of $\text{tr}(W_{\text{AV}} \Sigma^{-1})$ over $\Sigma \succ 0$ is*

$$\Sigma^*(\rho) = \begin{bmatrix} I_d & \rho I_d \\ \rho I_d & I_d \end{bmatrix}, \qquad \rho = \frac{\omega}{1 + \sqrt{1 - \omega^2}}. \tag{14}$$

The covariance and correlation formulas are derived in Appendix D.6.

The optimal distribution shape for multimodal tasks. Propositions 2 and 3 show that, under the total-variance constraint, $\mathcal{N}(0, \Sigma^*(\rho))$ uniquely minimizes $J_{W_{\mathrm{AV}}}(p_z)$; we adopt it as our target embedding distribution. The task-prior parameters π, τ^2_{ind} and τ^2_{same} are unknown before pretraining, so we retain this correlated Gaussian family and treat ρ as an empirical hyperparameter.

3.4 Enforcing the target distribution with SIGReg

Constraining the transformed embedding with SIGReg. SIGReg regularizes embeddings toward a standard Gaussian (Balestriero & LeCun, 2025). To adapt it to our correlated target, we first apply a fixed linear whitening transform that maps the target covariance to the identity (Kessy et al., 2018). We then apply SIGReg to the transformed embeddings. For the joint projection $\mathbf{z} = [\mathbf{e}_a; \mathbf{e}_v]$ and $|\rho| < 1$, the transform is

$$T_\rho \mathbf{z} = \left[\frac{\mathbf{e}_a + \mathbf{e}_v}{\sqrt{2(1+\rho)}} \; ; \; \frac{\mathbf{e}_a - \mathbf{e}_v}{\sqrt{2(1-\rho)}} \right]. \tag{15}$$

Since T_ρ is invertible, $T_\rho \mathbf{z} \sim \mathcal{N}(0, I_{2d})$ if and only if $\mathbf{z} \sim \mathcal{N}(0, \Sigma^*(\rho))$, so constraining the transformed embedding to a standard Gaussian enforces Eq. (14).

Our objective follows the prediction and regularization structure of LeJEPA. Following its view-counting convention, each audio-visual sample produces one global joint embedding $\mathbf{z}_g$ and V_l local embeddings $\mathbf{z}_1, \ldots, \mathbf{z}_{V_l}$, one per audio window or video frame. Here $\mathbf{z}_g, \mathbf{z}_i \in \mathbb{R}^{2d}$ share the block coordinates of $\mathbf{z} = [\mathbf{e}_a; \mathbf{e}_v]$. The prediction term is the coordinate-wise mean squared error:

$$\mathcal{L}_{\mathrm{pred}} = \frac{1}{2dV_l} \sum_{i=1}^{V_l} \|\mathbf{z}_i - \mathbf{z}_g\|^2 . \tag{16}$$

The distributional term $\mathcal{L}_{\mathrm{SIGReg}}$ applies SIGReg to the audio and video blocks of each embedding with the isotropic target, as in LeJEPA, and to the whitened local embeddings, with the isotropic target, corresponding to the correlated target before whitening. The overall objective is

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\mathrm{pred}} + \lambda \mathcal{L}_{\mathrm{SIGReg}}, \qquad \lambda \in [0, 1]. \tag{17}$$

4 Experiments

4.1 Experimental setup

We pretrain on AudioSet without using its class labels (Gemmeke et al., 2017). Because some videos are no longer available, our copies are smaller than the official lists; Appendix A gives the exact counts. We evaluate on VGGSound (VGG), a 309-class single-label classification benchmark (Chen et al., 2020a), and on the balanced and full training subsets of AudioSet (527 labels), denoted AudioSet-20K (AS-20K) and AudioSet-2M. We report top-1 accuracy for VGGSound and mean average precision (mAP) for AudioSet, in percent; A, V and A-V denote audio-only, video-only and audio-visual inputs.

Our encoder follows the ViT-B/16 configuration, with modality-specific patch embeddings and eleven modality-specific transformer blocks followed by a shared transformer block (Dosovitskiy et al., 2021; Gong et al., 2023). Audio is represented by a 128-bin log-mel spectrogram with 1,024 time positions, and video frames are resized to 224×224. A shared projection head maps the 768-dimensional backbone features to 128 dimensions. The reference configuration uses one jointly encoded global audio-visual view, seven local audio windows, and nine video-frame views sampled from a sixteen-frame pool. The encoder is initialized from ImageNet-MAE weights (He et al., 2022). Pretraining runs for 30 epochs with AdamW (Loshchilov & Hutter, 2019) at learning rate 2×10^{-4} and batch size 256, with $\lambda = 0.05$ and target correlation $\rho = 0.8$.

Table 1: Frozen-backbone probing: VGGSound top-1 accuracy and AudioSet-20K mAP (%, ↑). Attentive probing trains an attention-pooling head on the frozen encoder; linear probing trains a single linear layer on mean-pooled frozen features.

		Attentive probing						Linear probing					
		VGGSound			AudioSet-20K			VGGSound			AudioSet-20K		
Method	Backbone	A	V	A-V	A	V	A-V	A	V	A-V	A	V	A-V
CAV-MAE	ViT-B	40.7	30.2	46.1	33.3	18.8	36.6	39.2	30.8	43.8	32.7	18.1	35.1
EquiAV	ViT-B	41.2	31.6	46.3	33.5	19.9	36.6	41.3	31.7	46.0	32.2	18.7	34.9
AVSiam	ViT-B	36.9	32.7	43.8	29.6	21.7	34.0	37.6	**34.1**	42.9	30.7	**22.6**	34.1
CAV-MAE Sync	ViT-B	35.0	23.0	40.3	26.6	12.2	29.4	35.2	23.7	40.3	26.1	11.9	28.1
MM-LeJEPA	ViT-B	**43.4**	**33.8**	**48.6**	**35.8**	**22.2**	**37.6**	**42.3**	32.2	**46.9**	**33.4**	19.7	**36.3**

Table 2: Fine-tuning on VGGSound (top-1 accuracy) and AudioSet (mAP) (%, ↑). $*$: A and V from single-modality inference of the audio-visual model. ‡: supervised ImageNet-21K initialization.

		VGGSound			AudioSet-20K			AudioSet-2M		
Method	Backbone	A	V	A-V	A	V	A-V	A	V	A-V
MBT‡	ViT-B	52.3	51.2	64.1	31.3	27.7	43.9	41.5	31.3	49.6
AV-MAE	ViT-B	—	—	64.2	—	—	—	—	—	50.0
CAV-MAE	ViT-B	59.5	47.0	65.5	37.7	19.8	42.0	46.6	26.2	51.2
MAViL	ViT-B	60.8	50.9	67.1	41.8	24.8	44.9	48.7	30.3	53.3
EquiAV	ViT-B	61.0	50.7	67.1	42.4	25.7	46.6	49.1	30.1	54.6
AVSiam*‡	ViT-B	55.7	46.0	64.9	36.5	23.7	41.6	45.2	30.8	50.1
AV-JEPA*	ViT-B	—	—	57.1	23.7	10.3	29.6	26.0	12.8	32.7
MM-LeJEPA	ViT-B	60.5	49.5	66.9	38.9	21.7	43.1	46.2	27.3	51.1

4.2 Frozen-backbone probing

Attentive probing trains an attention-pooling head (Chen et al., 2024; Bardes et al., 2024) on the frozen encoder for 20 epochs, and linear probing trains a single linear layer on mean-pooled frozen features. Baselines use their released weights under the same code. Table 1 compares the frozen features. With attentive probing, MM-LeJEPA is best in all six cells, ahead of the strongest baseline in each column by 1.1–2.3 points on VGGSound and by 0.5–2.3 points on AudioSet-20K. With linear probing it is best in four of the six cells, including both audio-visual cells (46.9 vs. 46.0 and 36.3 vs. 35.1), while AVSiam, which shares one transformer across modalities, has the strongest video-only features. This is the setting the distributional target is designed for: the encoder is not adapted, so the scores reflect the embedding distribution itself.

4.3 Fine-tuning on three datasets

Fine-tuning attaches a classification head and uses AdamW (Loshchilov & Hutter, 2019) (learning rate 10^{-4}) for 10 epochs on VGGSound, Adam (Kingma & Ba, 2015) (learning rate 1.5×10^{-4}) for 15 on AudioSet-20K and Adam (learning rate 10^{-5}) for 15 on AudioSet-2M, with cosine schedules. Table 2 reports the fine-tuning results. MM-LeJEPA performs on par with the best open-source methods on all three datasets, without contrastive negatives, EMA teacher or self-training.

VGGSounder. VGGSounder (Zverev et al., 2025) re-annotates the VGGSound test set with multi-label, modality-specific labels. Following its protocol, we evaluate the VGGSound fine-tuned models on the background-music-free subset with the v1.0 labels, take the top-1 prediction, and report subset accuracy, F1 and hit rate separately for audio, visual and audio-visual inputs (Table 3; Appendix C reports the full test set). Baseline results are taken from the benchmark paper (Zverev et al., 2025) (arXiv v5, v1.0 labels). MM-LeJEPA obtains the highest score in all nine columns.

4.4 Sensitivity to target correlation, regularization weight and batch size

Table 6 varies the target correlation ρ. Moving it away from the reference degrades every score: mildly for $\rho = 0.2$ (by at most 0.7), clearly for $\rho = 0.99$ (by 2.1–4.1 points), and sharply for LeJEPA's isotropic target $\rho = 0$, whose attentive-probing scores fall by 8.8 and 9.8; Appendix B visualizes the effect of ρ on the embeddings. Tables 4 and 5 vary the regularization weight λ and

Table 3: VGGSounder results on the background-music-free subset (v1.0 labels, top-1, %, ↑). A, V and A-V denote audio, visual and audio-visual inputs.

	Subset accuracy			F1			Hit rate		
Method	A	V	A-V	A	V	A-V	A	V	A-V
CAV-MAE (Gong et al., 2023)	17.9	24.2	29.6	39.5	38.9	47.2	67.1	56.4	67.5
EquiAV (Kim et al., 2024)	15.8	13.6	24.4	33.7	22.9	39.3	57.1	33.3	56.2
AVSiam (Lin & Bertasius, 2024)	17.4	25.0	27.9	38.1	39.8	43.7	64.7	57.7	62.5
DeepAVFusion (Mo & Morgado, 2024)	13.9	13.8	26.2	29.1	23.3	41.8	49.4	33.8	59.8
MM-LeJEPA	**18.2**	**25.8**	**30.3**	**39.6**	**41.0**	**48.0**	**67.2**	**59.5**	**68.6**

Table 4: Regularization-weight study.

	Fine-tuning		Attentive probing	
λ	VGG	AS-20K	VGG	AS-20K
0.025	66.5	42.8	48.4	37.4
0.050	66.9	43.1	48.6	37.6
0.100	66.3	42.4	48.8	37.7
0.200	65.2	40.5	46.6	37.1

Table 5: Batch-size study.

	Fine-tuning		Attentive probing	
Batch	VGG	AS-20K	VGG	AS-20K
64	65.9	41.1	48.4	37.3
128	66.7	42.9	48.8	37.6
256	66.9	43.1	48.6	37.6
512	66.6	42.8	48.3	37.1

Table 6: Target-correlation study.

	Fine-tuning		Attentive probing	
ρ	VGG	AS-20K	VGG	AS-20K
0.00	64.6	31.5	39.8	27.8
0.20	66.2	42.9	48.1	37.0
0.40	66.5	43.0	48.2	37.1
0.60	66.7	42.9	48.5	37.4
0.80	66.9	43.1	48.6	37.6
0.99	64.8	39.0	45.1	34.4

Table 7: Component ablations.

	Fine-tuning		Attentive probing	
Setting	VGG	AS-20K	VGG	AS-20K
Full objective	66.9	43.1	48.6	37.6
Isotropic joint target ($\rho = 0$)	64.6	31.5	39.8	27.8
Mean and local joint pairs	66.5	43.0	47.9	37.3
Local joint pairs only	66.4	43.1	47.6	37.0
Without prediction	41.3	5.6	10.0	5.7
Without SIGReg	40.2	2.0	13.2	1.4

the batch size around the reference configuration. With $\lambda = 0.1$ the scores stay within 0.7 of the reference, whereas $\lambda = 0.2$ lowers three of the four scores by 1.7–2.6 points, so the prediction term must dominate the objective. Halving the batch to 128 changes the scores by at most 0.2 with no consistent sign, and doubling it to 512 lowers them by at most 0.5, so the scores are insensitive to the number of samples per SIGReg estimate in this range; only the smallest batch, 64, costs up to 2.0 on AudioSet-20K.

4.5 Component ablations

Table 7 varies where the joint regularizer is applied and removes each loss term. In the full objective the joint SIGReg term acts on the mean of the local embeddings, $\bar{\mathbf{z}} = \frac{1}{V_l} \sum_i \mathbf{z}_i$; "mean and local joint pairs" applies it to every local embedding as well, and "local joint pairs only" to the local embeddings alone. The per-embedding isotropic term is kept throughout.

Regularizing the local pairs, with or without the mean pair, changes the scores by at most one point. Removing either prediction or SIGReg collapses the representation, with all probing scores below 14, so both terms are necessary.

5 Conclusion

We presented MM-LeJEPA, a multimodal extension of LeJEPA that models independent-information and shared-content task components through their gradient moments. Under the stated correspondence and symmetry assumptions, minimizing the weighted Fisher information functional with a fixed total variance yields a correlated Gaussian target. A fixed whitening transform allows this target to be enforced with SIGReg while retaining an objective combining prediction and regularization. This enables multimodal pretraining without reconstruction decoders, contrastive negatives, or an EMA teacher. On public audio-visual classification benchmarks, MM-LeJEPA is competitive with open-source methods under both frozen probing and fine-tuning.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the grammar, clarity, and readability of text written by the authors. Generative AI was not used to generate research ideas, formulate hypotheses or claims, design the methodology or experiments, implement the method, analyze data, or interpret results. The authors reviewed all AI-assisted edits and take full responsibility for the final content of this work.

REFERENCES

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Relja Arandjelović and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017.

Relja Arandjelović and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision*, pp. 435–451, 2018.

Edson Araujo, Andrew Rouditchenko, Yuan Gong, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Leonid Karlinsky, Rogerio Feris, James R. Glass, and Hilde Kuehne. Cav-mae sync: Improving contrastive audio-visual mask autoencoders via fine-grained alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. SoundNet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 2022.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 2020b.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223, 2024.

Ciem Cornelissen, Sam Leroux, and Pieter Simoens. Le MuMo JEPA: Multi-modal self-supervised representation learning with learnable fusion tokens. *arXiv preprint arXiv:2603.24327*, 2026.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Harald Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

T. W. Epps and L. B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3015–3024, 2021.

Zhengcong Fei, Mingyuan Fan, and Junshi Huang. A-JEPA: Joint-embedding predictive architecture can listen. *arXiv preprint arXiv:2311.15830*, 2023.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 776–780, 2017.

Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Haoqi Fan, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, and Christoph Feichtenhofer. MAViL: Masked audio-video learners. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018.

Jongsuk Kim, Hyeongkeun Lee, Kyeongha Rho, Junmo Kim, and Joon Son Chung. EquiAV: Leveraging equivariance for audio-visual contrastive learning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Lukas Kuhn, Lucas Maes, Giuseppe Serra, Quentin Le Lidec, Yann LeCun, Randall Balestriero, and Florian Buettner. LeVJEPA: Efficient & scalable video pretraining without the heuristics. *arXiv preprint arXiv:2608.27395*, 2026.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview preprint, version 0.9.2, 2022.

Wenrui Li, Zhengyu Ma, Liang-Jian Deng, Hengyu Man, and Xiaopeng Fan. Modality-fusion spiking transformer network for audio-visual zero-shot learning. In *IEEE International Conference on Multimedia and Expo*, pp. 426–431, 2023.

Wenrui Li, Penghong Wang, Ruiqin Xiong, and Xiaopeng Fan. Spiking Tucker fusion transformer for audio-visual zero-shot learning. *IEEE Transactions on Image Processing*, 33:4840–4852, 2024.

Yan-Bo Lin and Gedas Bertasius. Siamese vision transformers are scalable audio-visual learners. In *European Conference on Computer Vision*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Shentong Mo and Pedro Morgado. Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12934–12945, 2021a.

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12475–12486, 2021b.

Elizbar A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision*, pp. 631–648, 2018.

Mandela Patrick, Yuki M. Asano, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

C. Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

Benjamin Robson, Santeri Mentu, Wenshuai Zhao, and Arno Solin. AV-JEPA: Extending LeJEPA to audio-visual self-supervised learning. *arXiv preprint arXiv:2607.15295*, 2026.

Revant Teotia, Adrien Bardes, Michael Rabbat, Sumit Chopra, Matthew Muckley, and Nicolas Ballas. Mjepa: A simple and scalable joint-embedding predictive architecture for audio-visual learning. In *European Conference on Computer Vision*, pp. 97–114. Springer, 2026.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Proceedings of the European Conference on Computer Vision*, 2020.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939, 2020.

Geoffrey S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320, 2021.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.

Daniil Zverev, Thaddäus Wiedemer, Ameya Prabhu, Matthias Bethge, Wieland Brendel, and A. Sophia Koepke. VGGSounder: Audio-visual evaluations for foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. Baseline numbers from arXiv:2508.08237v5 (June 2026).

APPENDIX

The appendix contains the following:

- Sizes of the datasets used, in Appendix A.
- Visualization of the effect of the target correlation on the embeddings, in Appendix B.
- VGGSounder results on the full test set, in Appendix C.
- Proof that the weighted Fisher information functional of Eq. (2) is a trace in Appendix D.1.
- Derivation of the two probe relations of Eq. (3), including their explicit constants, in Appendix D.2.
- Sources of the isotropic Fisher bound and its equality conditions in Appendix D.3.
- Proof of Proposition 1, Eqs. (11) and (12), in Appendix D.4.
- Proof of Proposition 2, Eq. (13), in Appendix D.5.
- Proof of Proposition 3, Eq. (14), in Appendix D.6.

A DATASET STATISTICS

Table 8 lists the number of clips in each dataset as used in this paper; some videos are no longer available, so our copies are smaller than the official lists. The pretraining and AudioSet-2M finetuning sets are the same 1,700,719 clips, which include all 18,020 clips of our AudioSet-20K copy.

Table 8: Number of clips in each dataset: official list versus the copy used in this paper.

Dataset	Official list	Used here
AudioSet training (balanced and unbalanced)	2,063,949	1,700,719
AudioSet-20K (balanced training)	22,160	18,020
AudioSet evaluation	20,371	16,621
VGGSound training	183,730	183,541
VGGSound test	15,446	15,428

B EFFECT OF THE TARGET CORRELATION ON THE EMBEDDINGS

Figure 2 shows the embeddings of two pretraining runs that differ only in the target correlation ρ, on 1,500 VGGSound test clips; no class labels are used. Each modality on its own (left two columns) looks alike under both settings, so the target correlation does not change the per-modality distribution. In the joint map (third column), the audio and video embeddings of the same clips occupy separate strands for $\rho = 0$ and share every small cluster for $\rho = 0.8$; the fraction of cross-modal nearest neighbors rises from 24% to 46% (50% for fully mixed modalities). The right column plots the paired coordinates $(e_{a,i}, e_{v,i})$ over all clips and coordinates, which is the quantity the target covariance constrains: the density is circular for $\rho = 0$ and elongated along the diagonal for $\rho = 0.8$, with a mean per-coordinate correlation of 0.66.

C VGGSOUNDER ON THE FULL TEST SET

Table 3 follows the benchmark protocol and evaluates on the background-music-free subset. Table 9 repeats the evaluation on the full VGGSounder test set (15,051 clips) with the same v1.0 labels, metrics and top-1 prediction. Baseline scores are computed with the same scorer from the per-clip predictions released by the benchmark.

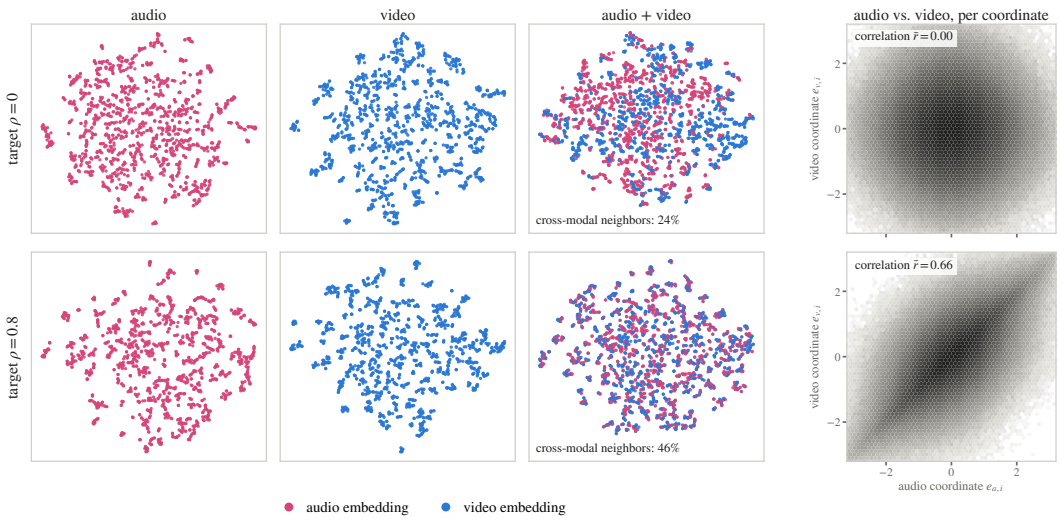

Figure 2: Embeddings of two pretraining runs that differ only in the target correlation ρ, on 1,500 VGGSound test clips (no class labels). Left three columns: t-SNE maps of the audio embeddings, of the video embeddings, and of both together (audio in magenta, video in blue); the number in the joint map is the fraction of each embedding's ten nearest neighbors, computed in the 128-dimensional embedding space, that belong to the other modality. Right column: density of the paired coordinates $(e_{a,i}, e_{v,i})$ over all clips and coordinates, with the mean per-coordinate correlation $\bar{r}$.

Table 9: VGGSounder results on the full test set (v1.0 labels, top-1, %, ↑). A, V and A-V denote audio, visual and audio-visual inputs, best per column in bold.

	Subset accuracy			F1			Hit rate		
Method	A	V	A-V	A	V	A-V	A	V	A-V
CAV-MAE (Gong et al., 2023)	16.7	24.2	29.4	38.1	38.5	46.7	65.1	55.2	66.0
EquiAV (Kim et al., 2024)	14.6	13.4	24.0	32.1	22.4	38.5	54.8	32.2	54.4
AVSiam (Lin & Bertasius, 2024)	16.1	24.7	27.2	36.6	39.0	42.7	62.6	55.9	60.4
DeepAVFusion (Mo & Morgado, 2024)	12.8	13.6	25.8	27.5	22.7	40.9	47.1	32.5	57.8
MM-LeJEPA	**16.9**	**25.8**	**29.9**	**38.1**	**40.5**	**47.3**	**65.1**	**58.1**	**66.9**

D PROOFS

D.1 TRACE REPRESENTATION OF THE WEIGHTED FISHER INFORMATION (EQ. 2)

The scalar identity $(\nabla\eta(\mathbf{q})^\top \nabla \log p_z(\mathbf{q}))^2 = \nabla \log p_z(\mathbf{q})^\top \nabla\eta(\mathbf{q}) \nabla\eta(\mathbf{q})^\top \nabla \log p_z(\mathbf{q})$ and $\mathbb{E}_\eta[\nabla\eta(\mathbf{q})\nabla\eta(\mathbf{q})^\top] = W$ give $\mathbb{E}_\eta[(\nabla\eta(\mathbf{q})^\top \nabla \log p_z(\mathbf{q}))^2] = \nabla \log p_z(\mathbf{q})^\top W \nabla \log p_z(\mathbf{q})$. Averaging over $\mathbf{q} \sim p_z$, we rewrite the quadratic form as a trace by the cyclicity of the trace and then move the expectation inside the trace, which is justified because the Fisher information is finite:

$$\begin{aligned} J_W(p_z) &= \mathbb{E}_\mathbf{q}[\nabla \log p_z(\mathbf{q})^\top W \nabla \log p_z(\mathbf{q})] \\ &= \mathbb{E}_\mathbf{q} \operatorname{tr}(W \nabla \log p_z(\mathbf{q}) \nabla \log p_z(\mathbf{q})^\top) = \operatorname{tr}(W \mathcal{I}(p_z)) \,. \end{aligned} \tag{18}$$

D.2 ROLE OF WEIGHTED FISHER INFORMATION IN THE TWO PROBES (EQ. 3)

Both relations follow the corresponding proofs of Balestriero & LeCun (2025) step by step (lemma and appendix numbers refer to arXiv v3). The task definitions, estimators and regularity conditions below are taken from LeJEPA and restated here to make the appendix self-contained. Our task prior changes only the term involving the second moment of the task gradient, and we expand that term alone. We write K for the embedding dimension ($K = 2d$ for the joint embedding) and $\mathcal{K}$ for the kernel.

Conditional on η, the training samples $\mathcal{D}$ are i.i.d. and independent of $\mathbf{q}$, and we define $\mathrm{Bias}(\mathbf{q}) = \mathbb{E}_{\mathcal{D}}[\widehat{y}(\mathbf{q}) \mid \eta, \mathbf{q}] - \eta(\mathbf{q})$ and $\mathrm{ISB} := \mathbb{E}_{\eta,\mathbf{q}}[\mathrm{Bias}(\mathbf{q})^2]$. Set zero-denominator predictions to zero; the remainder assumptions include this convention.

Radius-based k-NN probe: task and estimator. Following Balestriero & LeCun (2025, Appendix A.1), the embedding density p_z is C^3 with derivatives of order up to three bounded and with finite Fisher information and covariance. The unknown labels are given by a target function $\eta : \mathbb{R}^K \to \mathbb{R}$ assumed C^2; classification is handled by setting $\eta(\mathbf{z}) = \mathbb{P}(Y = 1 \mid \mathbf{z})$. The training data consist of the N embeddings together with their labels $\{(\mathbf{z}_n, \eta(\mathbf{z}_n))\}_{n=1}^N$, and we write $y_n := \eta(\mathbf{z}_n)$. The prediction at a query $\mathbf{q}$ averages the labels of the training points inside a ball of radius r_0,

$$\widehat{y}(\mathbf{q}) := \frac{1}{N_{r_0}(\mathbf{q})} \sum_{n\,:\,\|\mathbf{z}_n - \mathbf{q}\| \le r_0} y_n, \tag{19}$$

$$N_{r_0}(\mathbf{q}) := \#\left\{n : \|\mathbf{z}_n - \mathbf{q}\| \le r_0\right\}, \tag{20}$$

where the radius r_0 controls how many neighbors are averaged. Query points follow the training distribution, $\mathbf{q} \sim p_z$.

Following Lemma 4 of Balestriero & LeCun (2025), we write the bias expansion as

$$\mathrm{Bias}(\mathbf{q}) = \frac{r_0^2}{K+2}\left[\nabla\eta(\mathbf{q})^\top \nabla \log p_z(\mathbf{q}) + \tfrac{1}{2}\Delta\eta(\mathbf{q})\right] + \xi_{r_0}(\mathbf{q}). \tag{21}$$

The remainder ξ_{r_0} includes finite-sample corrections; we assume $\mathbb{E}_{\eta,\mathbf{q}}[\xi_{r_0}^2] = o(r_0^4)$ in the joint large-sample, small-radius limit. Following their notation we abbreviate $A(\mathbf{q}) := \nabla\eta(\mathbf{q})^\top \nabla \log p_z(\mathbf{q})$ and $C(\mathbf{q}) := \frac{1}{2}\Delta\eta(\mathbf{q})$. Squaring Eq. (21) and taking the expectation over $\mathbf{q} \sim p_z$ and over the task prior gives their three-term decomposition,

$$\mathbb{E}_{\eta,\mathbf{q}}\big[\mathrm{Bias}(\mathbf{q})^2\big] = \left(\frac{r_0^2}{K+2}\right)^2 \Big\{ \underbrace{\mathbb{E}_{\eta,\mathbf{q}}\big[A^2\big]}_{\text{score–gradient}} + 2\underbrace{\mathbb{E}_{\eta,\mathbf{q}}[AC]}_{\text{cross}} + \underbrace{\mathbb{E}_{\eta,\mathbf{q}}\big[C^2\big]}_{\text{curvature}} \Big\} + o(r_0^4). \tag{22}$$

Only the score–gradient term involves the second moment of the task gradient. Expanding this term with the joint prior of Section 3 in place of the isotropic one, and using the identity of Appendix D.1,

$$\begin{aligned} \mathbb{E}_{\eta,\mathbf{q}}\big[A^2\big] &= \mathbb{E}_{\mathbf{q}}\,\mathbb{E}_{\eta}\Big[\big(\nabla\eta(\mathbf{q})^\top \nabla \log p_z(\mathbf{q})\big)^2\Big] \\ &= \mathbb{E}_{\mathbf{q}}\,\mathrm{tr}\big(\nabla \log p_z(\mathbf{q}) \nabla \log p_z(\mathbf{q})^\top\, \mathbb{E}_{\eta}\big[\nabla\eta(\mathbf{q})\nabla\eta(\mathbf{q})^\top\big]\big) \\ &= \mathrm{tr}\big(W\,\mathbb{E}_{\mathbf{q}}\big[\nabla \log p_z(\mathbf{q}) \nabla \log p_z(\mathbf{q})^\top\big]\big) \\ &= \mathrm{tr}(W\mathcal{I}(p_z)) \;=\; J_W(p_z). \end{aligned} \tag{23}$$

This modifies the score–gradient term. Where Balestriero & LeCun (2025) substitute $\mathbb{E}_\eta[\nabla\eta\nabla\eta^\top] = \tau_g^2 I_K$ at this point and obtain $\tau_g^2 J(p_z)$, the structured prior retains the matrix W and yields $J_W(p_z)$; setting $W = \tau_g^2 I_K$ recovers their expression exactly.

For the cross and curvature terms we additionally assume $\mathbb{E}_\eta[\nabla\eta(\mathbf{q})\,\Delta\eta(\mathbf{q})] = 0$ and $\mathbb{E}_\eta[(\Delta\eta(\mathbf{q}))^2] = \kappa^2$, where $\kappa^2 < \infty$ is independent of $\mathbf{q}$ and p_z. We retain the cross and curvature decomposition of Balestriero & LeCun (2025, Appendix B.4) under these additional moment assumptions. The cross term vanishes under the assumption $\mathbb{E}_\eta[\nabla\eta\,\Delta\eta] = 0$ stated above. The curvature term equals $\frac{1}{4}\mathbb{E}_\eta[(\Delta\eta)^2] = \kappa^2/4$, which does not depend on p_z; multiplied by the prefactor it contributes at order $O(r_0^4)$, the same order as the J_W term, and is therefore a constant offset rather than a vanishing remainder. Substituting the three terms into Eq. (22),

$$\mathbb{E}_{\eta,\mathbf{q}}\big[\mathrm{Bias}(\mathbf{q})^2\big] = \frac{r_0^4}{(K+2)^2}\, J_W(p_z) + \frac{r_0^4\kappa^2}{4(K+2)^2} + o(r_0^4), \tag{24}$$

which is the first relation of Eq. (3), with $c_{k\mathrm{NN},0} = r_0^4\kappa^2/(4(K+2)^2)$ from the curvature term, $c_{k\mathrm{NN},1} = r_0^4(K+2)^{-2}$, and $\varepsilon_{k\mathrm{NN}} = o(r_0^4)$ from the remainder (the cross term vanishes). As in Balestriero & LeCun (2025), what this relation establishes is not that the integrated squared bias equals the functional, but that the part of it which varies with the shape of p_z at leading order is $J_W(p_z)$.

Nadaraya–Watson kernel probe: task and estimator. Following Balestriero & LeCun (2025, Appendix A.2), consider a nonnegative kernel $\mathcal{K} : \mathbb{R}^K \to \mathbb{R}$ with the standard properties

$$\int_{\mathbb{R}^K} \mathcal{K}(\mathbf{u})\,\mathrm{d}\mathbf{u} = 1, \qquad \text{(normalized)} \tag{25}$$

$$\int_{\mathbb{R}^K} \mathbf{u}\,\mathcal{K}(\mathbf{u})\,\mathrm{d}\mathbf{u} = 0, \qquad \text{(zero first moment)} \tag{26}$$

$$\int_{\mathbb{R}^K} \mathbf{u}\mathbf{u}^\top \mathcal{K}(\mathbf{u})\,\mathrm{d}\mathbf{u} = \mu_2(\mathcal{K}) I_K, \qquad \text{(isotropic second moment)} \tag{27}$$

$$\int_{\mathbb{R}^K} \mathcal{K}(\mathbf{u})^2\,\mathrm{d}\mathbf{u} < \infty, \qquad \text{(finite roughness)} \tag{28}$$

for some $\mu_2(\mathcal{K}) \in (0, \infty)$. For a bandwidth $h > 0$ write $\mathcal{K}_h(\mathbf{t}) := h^{-K}\mathcal{K}(\mathbf{t}/h)$. The Nadaraya–Watson estimator at a query $\mathbf{q}$ is

$$\widehat{y}(\mathbf{q}) := \frac{\sum_{n=1}^N \mathcal{K}_h(\mathbf{q} - \mathbf{z}_n)\, y_n}{\sum_{n=1}^N \mathcal{K}_h(\mathbf{q} - \mathbf{z}_n)}. \tag{29}$$

The targets are given by $\eta(\mathbf{z}_n) = \mathbb{E}[Y_n \mid \mathbf{z}_n]$, so the task function is the same as that used for the k-NN probe, and the conditional variance of the target is denoted by $\sigma^2(\mathbf{z}) := \mathrm{Var}(Y \mid \mathbf{z})$. The regularity conditions of the k-NN derivation are retained, and p_z is additionally assumed to have light enough tails that $p_z(\mathbf{z}) \to 0$ and $z_j\, p_z(\mathbf{z}) \to 0$ as $\|\mathbf{z}\| \to \infty$ for every coordinate j. The task prior is supported on the smoothness class

$$\mathcal{M}(L, B) := \left\{\eta \in C^2(\mathbb{R}^K) : \|\nabla\eta(\mathbf{q})\| \le L,\ |\Delta\eta(\mathbf{q})| \le B,\ \forall \mathbf{q} \in \mathbb{R}^K\right\}. \tag{30}$$

Following Lemma 5 of Balestriero & LeCun (2025), for $h \to 0$ and $Nh^K \to \infty$ we write

$$\mathrm{Bias}(\mathbf{q}) = \frac{h^2\mu_2(\mathcal{K})}{2}\left[\Delta\eta(\mathbf{q}) + 2\nabla\eta(\mathbf{q})^\top \nabla \log p_z(\mathbf{q})\right] + \xi_h(\mathbf{q}). \tag{31}$$

The remainder ξ_h includes finite-sample corrections; we assume $\mathbb{E}_{\eta,\mathbf{q}}[\xi_h^2] = o(h^4)$ in the joint large-sample, small-bandwidth limit.

We now follow the steps of Balestriero & LeCun (2025, Appendix B.7). Squaring Eq. (31), averaging over tasks and queries $\mathbf{q} \sim p_z$ and applying $(a+b)^2 \le 2a^2 + 2b^2$ pointwise,

$$\begin{aligned}
\mathbb{E}_{\eta,\mathbf{q}}\big[\mathrm{Bias}(\mathbf{q})^2\big] &= \left(\frac{h^2\mu_2(\mathcal{K})}{2}\right)^2 \mathbb{E}_\eta \int \left[\Delta\eta + 2\nabla\eta^\top\nabla\log p_z\right]^2 p_z + o(h^4) \\
&\le \left(\frac{h^2\mu_2(\mathcal{K})}{2}\right)^2 \left[2\mathbb{E}_\eta \int (\Delta\eta)^2\, p_z + 8\mathbb{E}_\eta \int \left(\nabla\eta^\top\nabla\log p_z\right)^2 p_z\right] + o(h^4).
\end{aligned} \tag{32}$$

The first integral is bounded exactly as in LeJEPA's proof: the curvature bound of Eq. (30) gives $\int(\Delta\eta)^2 p_z \le B^2$.

We modify only the treatment of the second integral. Balestriero & LeCun (2025) bound it with the worst-case gradient norm of the class $\mathcal{M}(L, B)$, applying Cauchy–Schwarz and $\|\nabla\eta\| \le L$ to obtain $L^2 J(p_z)$, which does not retain the directional structure of the task prior. We instead take the task expectation first, so that the task-moment identity in Appendix D.1 applies pointwise and the integral is evaluated rather than bounded,

$$\begin{aligned}
\int \mathbb{E}_\eta\Big[\left(\nabla\eta(\mathbf{q})^\top\nabla\log p_z(\mathbf{q})\right)^2\Big] p_z(\mathbf{q})\,\mathrm{d}\mathbf{q} &= \int \nabla\log p_z(\mathbf{q})^\top W \nabla\log p_z(\mathbf{q})\, p_z(\mathbf{q})\,\mathrm{d}\mathbf{q} \\
&= \mathrm{tr}(W\mathcal{I}(p_z)) \;=\; J_W(p_z).
\end{aligned} \tag{33}$$

Substituting both integrals into Eq. (32),

$$\mathbb{E}_{\eta,\mathbf{q}}\big[\mathrm{Bias}(\mathbf{q})^2\big] \le \left(\frac{h^2\mu_2(\mathcal{K})}{2}\right)^2 \left[2B^2 + 8J_W(p_z)\right] + o(h^4), \tag{34}$$

which is the second relation of Eq. (3), with

$$c_{\mathrm{ker},0} = 2B^2\left(\frac{h^2\mu_2(\mathcal{K})}{2}\right)^2, \qquad c_{\mathrm{ker},1} = 8\left(\frac{h^2\mu_2(\mathcal{K})}{2}\right)^2, \qquad \varepsilon_{\mathrm{ker}} = o(h^4). \tag{35}$$

These task-averaged bias relations motivate the use of $J_W(p_z)$ as the distribution-design criterion in Section 3.

D.3 Sources of the isotropic Fisher bound and equality conditions (Eq. 4)

The unweighted Fisher bound and Gaussian equality condition used in Eq. (4) are established in Lemma 6 of Balestriero & LeCun (2025, Appendix B.4, p. 30); the trace-constraint step and its equality condition are given in Step 2(a) of the same appendix. The resulting isotropic Gaussian optimum $p_z^* = \mathcal{N}(0, I_K)$ is the trace-constrained case of that appendix's concluding optimality statement.

D.4 Proof of Proposition 1 (Eqs. 11 and 12)

Proof of Proposition 1. By the independent-information conditions of Assumption 2, W_{ind} has diagonal blocks $\tau_{\text{ind}}^2 I_d$ and zero cross blocks. For the shared-content component, Eq. (8) gives identical gradients on the two sides, so all four blocks of W_{same} equal $\tau_{\text{same}}^2 I_d$. This proves Eq. (11). Component independence and zero means give $\mathbb{E}_\eta[\nabla\eta_{\text{ind}}\nabla\eta_{\text{same}}^\top] = 0$, and expanding Eq. (6) yields

$$\begin{aligned}\mathbb{E}_\eta[\nabla\eta\nabla\eta^\top] &= (1-\pi)W_{\text{ind}} + \pi W_{\text{same}} \\ &\quad + \sqrt{\pi(1-\pi)}\,\mathbb{E}_\eta[\nabla\eta_{\text{ind}}\nabla\eta_{\text{same}}^\top + \nabla\eta_{\text{same}}\nabla\eta_{\text{ind}}^\top] \\ &= (1-\pi)W_{\text{ind}} + \pi W_{\text{same}}.\end{aligned}$$

The weights $(1-\pi)$ and π in Eq. (12) are thus the squared component amplitudes of Eq. (6). The combined task prior also satisfies $\mathbb{E}_\eta[\nabla\eta] = 0$. Since $\tau_{\text{ind}}^2 > 0$ and $\pi < 1$, we have $1-\omega > 0$, so the two eigenvalues $\tau_{\text{AV}}^2(1+\omega)$ and $\tau_{\text{AV}}^2(1-\omega)$ of W_{AV} (Eq. (40)) are both positive, *i.e.* $W_{\text{AV}} \succ 0$. □

D.5 Proof of Proposition 2 (Eq. 13)

Proof of Proposition 2. Let p_z satisfy the conditions of Proposition 2, namely zero mean, covariance $\Sigma = \mathbb{E}_{\mathbf{z}\sim p_z}[\mathbf{z}\mathbf{z}^\top] \succ 0$, $p_z \in C^3$ with bounded derivatives up to order three, finite Fisher information and covariance, and p_z everywhere positive on $\mathbb{R}^{2d}$. Under these conditions, the proof of Lemma 6 of LeJEPA (Balestriero & LeCun, 2025, Appendix B.4) applies the matrix Cramér–Rao bound (Cramér, 1946; Rao, 1945) to the family of shifted densities $p_{\boldsymbol{\mu}}(\mathbf{z}) = p_z(\mathbf{z} - \boldsymbol{\mu})$, whose Fisher information with respect to the shift $\boldsymbol{\mu}$ is $\mathcal{I}(p_z)$. It yields $\mathcal{I}(p_z) \succeq \Sigma^{-1}$, with equality if and only if $p_z = \mathcal{N}(0, \Sigma)$.

Multiplying $\mathcal{I}(p_z) - \Sigma^{-1} \succeq 0$ by $W_{\text{AV}}^{1/2}$ on both sides and taking traces,

$$J_{W_{\text{AV}}}(p_z) - \operatorname{tr}\big(W_{\text{AV}}\Sigma^{-1}\big) = \operatorname{tr}\Big(W_{\text{AV}}^{1/2}\left(\mathcal{I}(p_z) - \Sigma^{-1}\right)W_{\text{AV}}^{1/2}\Big) \;\geq\; 0, \tag{36}$$

which is Eq. (13). If Eq. (13) holds with equality, the positive semidefinite matrix $W_{\text{AV}}^{1/2}\left(\mathcal{I}(p_z) - \Sigma^{-1}\right)W_{\text{AV}}^{1/2}$ has zero trace and is therefore 0. Since $W_{\text{AV}} \succ 0$, the square root $W_{\text{AV}}^{1/2}$ is invertible, so $\mathcal{I}(p_z) = \Sigma^{-1}$ and the equality condition above gives $p_z = \mathcal{N}(0, \Sigma)$. Conversely, if $p_z = \mathcal{N}(0, \Sigma)$ then $\mathcal{I}(p_z) = \Sigma^{-1}$ and Eq. (13) holds with equality. □

D.6 Proof of Proposition 3 (Eq. 14)

Proof of Proposition 3. The proof has two steps. *Step 1 (a lower bound from the Cauchy–Schwarz inequality).* In the Frobenius inner product $\langle A, B\rangle_F = \operatorname{tr}(A^\top B)$, take $A = W_{\text{AV}}^{1/2}\Sigma^{-1/2}$ and $B = \Sigma^{1/2}$. By the cyclic property of the trace, $\langle A, B\rangle_F = \operatorname{tr}(\Sigma^{-1/2}W_{\text{AV}}^{1/2}\Sigma^{1/2}) = \operatorname{tr}(W_{\text{AV}}^{1/2})$. We also have $\|A\|_F^2 = \operatorname{tr}(W_{\text{AV}}\Sigma^{-1})$ and $\|B\|_F^2 = \operatorname{tr}(\Sigma) = 2d$ by the trace constraint, so that

$$\Big[\operatorname{tr}\Big(W_{\text{AV}}^{1/2}\Big)\Big]^2 \leq \operatorname{tr}\big(W_{\text{AV}}\Sigma^{-1}\big)\operatorname{tr}(\Sigma), \qquad \text{that is} \qquad \operatorname{tr}\big(W_{\text{AV}}\Sigma^{-1}\big) \;\geq\; \frac{\Big[\operatorname{tr}\Big(W_{\text{AV}}^{1/2}\Big)\Big]^2}{2d}. \tag{37}$$

Equality holds if and only if A and B are proportional, *i.e.* $W_{\mathrm{AV}}^{1/2}\Sigma^{-1/2} = c\,\Sigma^{1/2}$, that is $\Sigma \propto W_{\mathrm{AV}}^{1/2}$; the trace constraint $\mathrm{tr}(\Sigma) = 2d$ then fixes the constant,

$$\Sigma^* = \frac{2d}{\mathrm{tr}\Big(W_{\mathrm{AV}}^{1/2}\Big)}\, W_{\mathrm{AV}}^{1/2}. \tag{38}$$

Step 2 (computing $W_{\mathrm{AV}}^{1/2}$ and normalizing). We first diagonalize W_{AV}: let

$$U = \frac{1}{\sqrt{2}}\begin{bmatrix} I_d & I_d \\ I_d & -I_d \end{bmatrix}, \qquad U^\top U = I_{2d}, \tag{39}$$

which replaces $\mathbf{z} = [\mathbf{e}_a; \mathbf{e}_v]$ by $\big[(\mathbf{e}_a + \mathbf{e}_v)/\sqrt{2}\,;\, (\mathbf{e}_a - \mathbf{e}_v)/\sqrt{2}\big]$. A direct computation gives

$$U^\top W_{\mathrm{AV}} U = \tau_{\mathrm{AV}}^2 \begin{bmatrix} (1+\omega) I_d & 0 \\ 0 & (1-\omega) I_d \end{bmatrix}, \tag{40}$$

so U diagonalizes W_{AV}, the two eigenvalues $\tau_{\mathrm{AV}}^2(1+\omega)$ and $\tau_{\mathrm{AV}}^2(1-\omega)$ each having multiplicity d. Taking the square root in this basis and transforming back to the original basis,

$$\begin{aligned} W_{\mathrm{AV}}^{1/2} &= U\,\tau_{\mathrm{AV}} \begin{bmatrix} \sqrt{1+\omega}\, I_d & 0 \\ 0 & \sqrt{1-\omega}\, I_d \end{bmatrix} U^\top \\ &= \frac{\tau_{\mathrm{AV}}}{2} \begin{bmatrix} \big(\sqrt{1+\omega} + \sqrt{1-\omega}\big)\, I_d & \big(\sqrt{1+\omega} - \sqrt{1-\omega}\big)\, I_d \\ \big(\sqrt{1+\omega} - \sqrt{1-\omega}\big)\, I_d & \big(\sqrt{1+\omega} + \sqrt{1-\omega}\big)\, I_d \end{bmatrix}. \end{aligned} \tag{41}$$

This can be checked by squaring: the diagonal blocks give $\tau_{\mathrm{AV}}^2[(1+\omega) + (1-\omega)]/2 = \tau_{\mathrm{AV}}^2$ and the off-diagonal blocks give $\tau_{\mathrm{AV}}^2[(1+\omega) - (1-\omega)]/2 = \tau_{\mathrm{AV}}^2\omega$. Its trace is $\mathrm{tr}(W_{\mathrm{AV}}^{1/2}) = d\,\tau_{\mathrm{AV}}\big(\sqrt{1+\omega} + \sqrt{1-\omega}\big)$, so τ_{AV} and d cancel in Eq. (38). Multiplying the numerator and denominator of ρ by $\sqrt{1+\omega} + \sqrt{1-\omega}$ then gives

$$\Sigma^*(\rho) = \begin{bmatrix} I_d & \rho I_d \\ \rho I_d & I_d \end{bmatrix}, \qquad \rho = \frac{\sqrt{1+\omega} - \sqrt{1-\omega}}{\sqrt{1+\omega} + \sqrt{1-\omega}} = \frac{\omega}{1 + \sqrt{1-\omega^2}}. \tag{42}$$

Both diagonal blocks $\Sigma_{aa}^*, \Sigma_{vv}^*$ of Eq. (42) are I_d, so $\mathrm{tr}(\Sigma_{aa}^*) = \mathrm{tr}(\Sigma_{vv}^*) = d$ follows directly: equal total variances for the two modalities are a consequence of this optimum, not an additional per-modality variance constraint. □